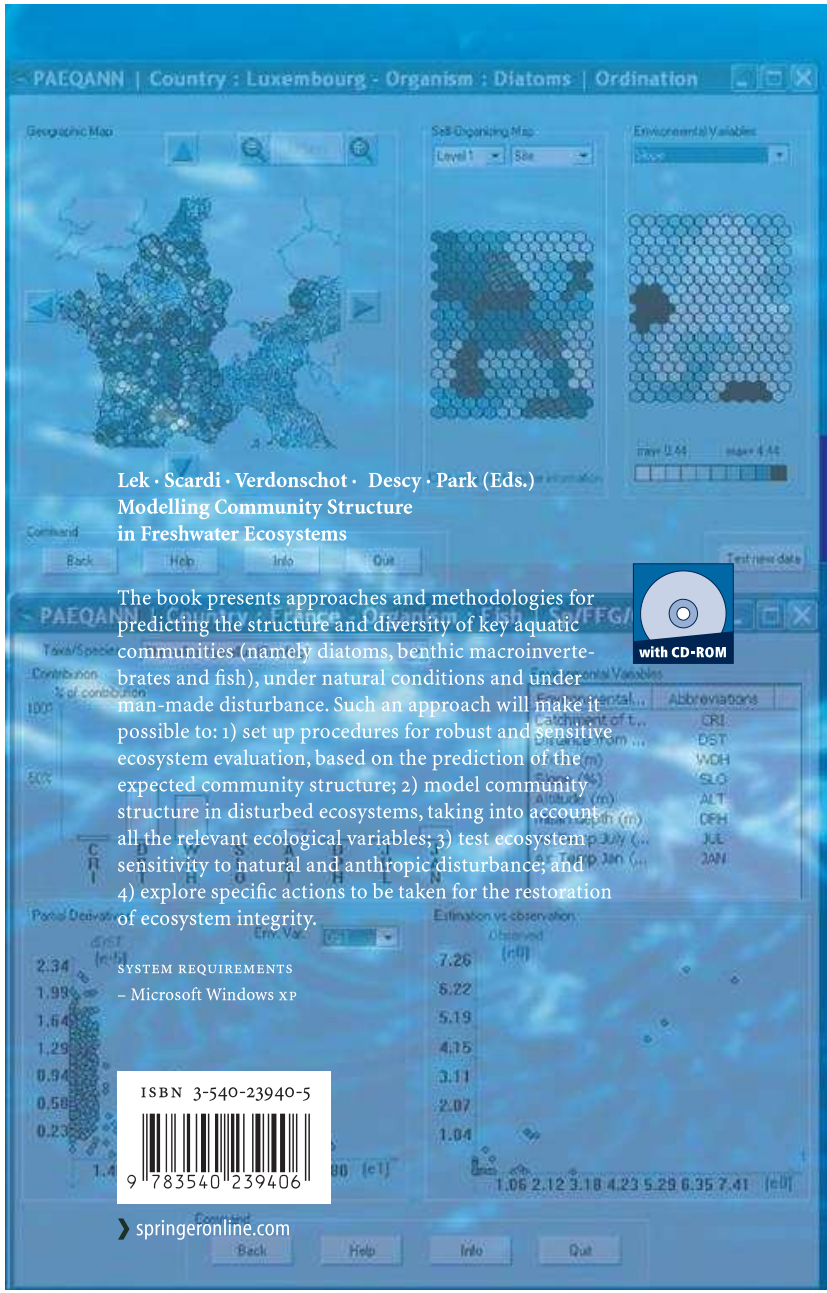


Lek · Scardi · Verdonschot · Descy · Park (Eds.)
Modelling Community Structure in Freshwater Ecosystems

The book presents approaches and methodologies for predicting the structure and diversity of key aquatic communities (namely diatoms, benthic macroinvertebrates and fish), under natural conditions and under man-made disturbance. Such an approach will make it possible to: 1) set up procedures for robust and sensitive ecosystem evaluation, based on the prediction of the expected community structure; 2) model community structure in disturbed ecosystems, taking into account all the relevant ecological variables; 3) test ecosystem sensitivity to natural and anthropic disturbance; and 4) explore specific actions to be taken for the restoration of ecosystem integrity.

with CD-ROM

SYSTEM REQUIREMENTS
– Microsoft Windows XP

ISBN 3-540-23940-5
9 783540 239406

springeronline.com

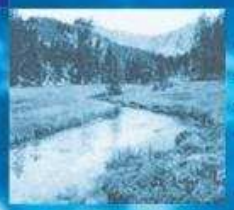
Lek · Scardi · Verdonschot
Descy · Park (Eds.)

Modelling Community Structure in Freshwater Ecosystems

with CD-ROM

Sovan Lek
Michele Scardi
Piet F.M. Verdonschot
Jean-Pierre Descy
Young-Seuk Park
Editors

Modelling Community Structure in Freshwater Ecosystems



with CD-ROM

Springer

3.8 Optimisation of artificial neural networks for predicting fish assemblages in rivers*

Scardi M†, Cataudella S, Ciccotti E, Di Dato P, Maio G, Marconato E, Salviati S, Tancioni L, Turin P, Zanetti M

Introduction

Fish assemblages are among the most sensitive and reliable indicators of the ecological status of stream and rivers (Fausch et al. 1990). Fish assemblages are able to integrate over both time and space the biological response to ecological processes more effectively than other biotic components (Harris 1995). Sampling fish fauna, of course, is not as simple as sampling other organisms, but in spite of this problem indices of biotic integrity based on fish have been developed and are now widely accepted (Karr 1981; Karr et al. 1986). Targeting fish fauna in environmental monitoring activities is effective not only from the ecological point of view, but also in the light of the need for straightforward communication with decision-makers as well as with other stakeholders. In fact, fish are probably the most direct and intuitive expression of aquatic ecosystem quality (McCormick et al. 2000).

Therefore, it is not surprising that composition, abundance and age structure of fish fauna are considered as some of the main biological quality elements for the classification of the ecological status of surface water in the EU Water Framework Directive (i.e. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy).

This Directive also states that biological reference conditions have to be established for each type of water body. These reference conditions are based on community structure and take into account all the biological quality elements, thus including fish fauna as well as benthic macroinvertebrates and aquatic flora. Hence, modeling fish assemblage composition on the basis of biotic and abiotic environmental descriptors will play a major role in the implementation of the Water Framework Directive and, more generally, in the management of aquatic ecosystems.

Predicting fish fauna as well as other biotic assemblages is not only relevant to the definition of reference conditions that are aimed at the evaluation of environmental quality. In fact, it is also an important achievement in scientific research, e.g. as a framework for studies on species interactions, and it can be very useful for a number of other applied tasks. In particular, species composition models may support environmental management by simulating different environmental scenarios and pointing out the most critical factors that need changes or regulation. Sensitivity analyses of the species composition models play a relevant role in this kind of studies.

* This chapter has been supported by the EU 5th Framework Programme PAEQANN project ["Predicting Aquatic Ecosystem Quality using Artificial Neural Networks: impact of environmental characteristics on the structure of aquatic communities (algae, benthic and fish fauna)", URL: http://aquaeco.ups-tlse.fr/], under contract EVK1-CT1999-00026.

† Correspondence: mscardi@mclink.it

Even though the idea of modeling fish fauna composition on the basis of environmental variables is not new (e.g. Faush et al. 1988), only recently Artificial Neural Networks (ANNs) have been applied to this problem. ANNs have been used to predict fish species richness (e.g. Guegan et al. 1998) as well as density and biomass of single fish populations (Baran et al. 1996; Lek et al. 1996a,b; Mastrorillo et al. 1997b) and ecological characteristics of fish assemblages (Aguilar Ibarra et al. 2003). As far as fish assemblage composition at the river basin scale is considered, only a few models have been developed so far, either using conventional statistical methods (e.g. Oberdorff et al. 2001) or ANNs (Boët and Fhus 2000; Joy and Death, # 3.5; Olden and Jackson 2001). A very useful introduction to the ecological applications of ANNs can be found in Lek and Guégan (1999).

ANNs and other modelling techniques that have been developed and formerly applied in other disciplines have often been introduced into ecological applications with no modification. In most cases this was not a problem and very useful results were obtained anyway. However, in ecological modelling adaptations of the modelling techniques are sometimes required in order to fit particular needs or to properly exploit the available information. This is certainly the case of species composition models, as the data that are involved in this kind of application cannot be regarded as mere numbers, because each species has a different ecological "meaning", which in turn depends on its coenotic context.

This chapter will present a case study about fish assemblages from some river basins in north-eastern Italy, showing how the above-mentioned problem can be tackled by developing ecologically enhanced ANNs.

Data set

The ANN models presented in this study are based on a data set that included sampling sites from several river basins in the Veneto region (north-eastern Italy), as shown in Fig. 3.8.1. The data set consisted of 264 records and it comprised two groups of variables. The first group included the variables to be predicted by the models, i.e. 34 fish species, whereas the second group embraced 20 predictive environmental variables, as shown in Tables 3.8.1 and 3.8.2 respectively.

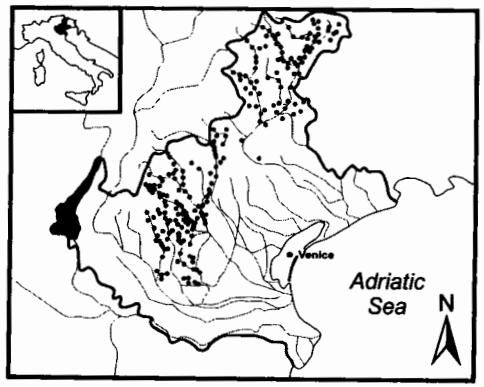


Figure 3.8.1 The sampling sites (black dots) were located in several river basins in the Veneto region (NE Italy).

Fish were collected by means of electrofishing gear. Either direct current or pulsed direct current electrofishing devices were used in streams and small rivers, while these tools were supported by nets when only part of larger rivers was sampled. Basically, in the latter case the electrofishing area was closed by means of nets that also acted as a sampling device.

Table 3.8.1 List of the fish species in the Veneto data set. Modeled species are on white background, while species that were excluded (see text) are on grey background. Italian names are shown in parentheses for those species that do not have an English name.

N	Scientific name	English name
1	*Salmo (trutta) trutta (Linnaeus 1758)*	Sea Trout
2	*Leuciscus cephalus (Linnaeus 1758)*	Chub
3	*Padogobius martensii (Günther 1861)*	(Ghiozzo di fiume)
4	*Scardinius erythrophthalmus (Linnaeus 1758)*	Rudd
5	*Esox lucius (Linnaeus 1758)*	European Pike
6	*Rutilus erythrophthalmus (Zerunian 1982)*	(Triotto)
7	*Alburnus alburnus alborella (De Filippi 1844)*	Bleak
8	*Cottus gobio (Linnaeus 1756)*	Bullhead
9	*Tinca tinca (Linnaeus 1758)*	Tench
10	*Cobitis taenia (Linnaeus 1758)*	Spined loach
11	*Phoxinus phoxinus (Linnaeus 1758)*	Minnow
12	*Anguilla anguilla (Linnaeus 1758)*	European Eel
13	*Knipowitschia punctatissima (Canestrini 1864)*	(Panzarolo)
14	*Salmo (trutta) marmoratus (Cuvier 1817)*	Marble Trout
15	*Sabanejewia larvata (DeFilippi 1859)*	Italian Loach
16	*Ictalurus melas (Rafinesque 1820)*	Black Bullhead
17	*Lepomis gibbosus (Linnaeus 1758)*	Pumpkinseed
18	*Barbus plebejus (Bonaparte 1839)*	Italian Barbel
19	*Chondrostoma genei (Bonaparte 1839)*	South Europe Nase
20	*Gasterosteus aculeatus (Linnaeus 1758)*	Three-spined Stickleback
21	*Carassius auratus (Linnaeus 1758)*	Crucian Carp
22	*Gobio gobio (Linnaeus 1758)*	Gudgeon
23	*Leuciscus souffia (Risso 1826)*	Blageon
24	*Thymallus thymallus (Linnaeus 1758)*	Grayling
25	*Lampetra zanandreai (Vladykov 1955)*	Po Brook Lamprey
26	*Gambusia holbrooki (Girard 1859)*	Eastern mosquitofish
27	*Barbus meridionalis*	Meriditerranean Barbel
28	*Micropterus salmoides (Lacepede 1802)*	Large-Mouthed Bass
29	*Perca fluviatilis (Linnaeus 1758)*	Perch
30	*Abramis brama (Linnaeus 1758)*	Common Bream
31	*Cyprinus carpio (Linnaeus 1758)*	Common Carp
32	*Salvelinus fontinalis M.*	Brook Char
33	*Oncorhynchus mykiss (Walbaum 1792)*	Rainbow Trout
34	*Salmo (trutta) hybr. trutta/marmoratus*	Sea Trout-Marble Trout hybrid

Two fish taxa, namely *Oncorhynchus mykiss*, i.e. the rainbow trout, and *Salmo (trutta)* hybr. *trutta/marmoratus*, i.e. a sea trout - marble trout hybrid (on grey background in Table 3.8.1), were excluded from the models, as their distribution only partly depends on environmental variables. In fact, the distribution of the first taxon is linked to the artificial release of reared juveniles, while that of the second taxon is clearly not independent of the

distribution of the two parent species and is probably associated to problems in species identification too.

Some of the available records refer to sampling activities that were carried out at the same site at two different times, thus representing the local interannual variability of both the fish fauna and the environmental variables.

The fish fauna composition was described using binary variables, i.e. presence or absence of each taxon. Quantitative data, although available in most cases, were not considered for model development as they were not sufficiently accurate because of the combined effects of varying efficiency of the electrofishing gear and morphodynamic heterogeneity of the sampling sites. The environmental variables were coded in different ways, either as quantitative or semi-quantitative data, and all the non-binary variables were normalized by rescaling them in the [0,1] interval.

Table 3.8.2 Environmental descriptors used as input (i.e. predictive) variables in the models.

1	elevation (m)
2	mean depth (m)
3	runs (area, %)
4	pools (area, %)
5	riffles (area, %)
6	mean width (m)
7	boulders (area, %)
8	rocks and pebbles (area, %)
9	gravel (area, %)
10	sand (area, %)
11	silt and clay (area, %)
12	stream velocity (score, 0-5)
13	vegetation covering (area, %)
14	shade (%)
15	anthropogenic disturbance (score, 0-4)
16	pH
17	conductivity ($\mu S\ cm^{-1}$)
18	gradient (%)
19	catchment area (km^2)
20	distance from source (km)

The whole data set was divided into three subsets for training, validating and testing the ANN models. The training data set included 50% of the records (n=132), whereas the validation and the test data sets included 25% each (n=66). Every record was assigned to a different subset after sorting all the records according to the elevation of the sampling sites. Starting from the highest elevation, the records were divided into the above-mentioned subsets by assigning uneven records to the training subset and by assigning each couple of successive even records to the validation and test subset, respectively. This way, the records in each group of four were assigned to the ($1x$) training, ($2x$) validation, ($3x$) training and ($4x$) test data subset, with x ranging from 1 to 66. This break up strategy allowed a homogeneous allocation of records for different elevations classes among the three subsets, thus stratifying the procedure on the basis of the most relevant environmental variable.

Neural network training

The most common type of ANN, i.e. the multilayer perceptron, was used for modeling the fish fauna composition. The error back-propagation algorithm (Rumelhart et al. 1986a) was used for training the ANNs, both in its original formulation and in a modified version that will be described later in this chapter. Other training algorithms were not tested because the theoretical advantages they might provide (e.g. quicker training) are not really relevant for ecological applications.

ANNs with 20 input nodes, 32 output nodes and 17 nodes in the hidden layer were selected after a set of empirical tests involving ANNs with different numbers of nodes in the hidden layer (from 10 to 40 nodes). The architecture selected was the one that provided the minimum overall error with respect to an independent test set. However, the selection of the number of nodes in the hidden layer was not a critical issue, as the differences among the models were negligible. Sigmoid activation functions [i.e. $f(x)=1/(1-e^{-x})$] were used both in the hidden and in the output nodes of all the ANNs that have been trained and used in this study.

In order to prevent overtraining, i.e. to avoid that the ANN "learned by heart" the fish fauna composition at each known site while losing its generalization ability, different strategies were adopted. The first strategy involved stopping the training procedure early. In other words, the training procedure was terminated as soon as the error, computed on the basis of the validation set only, ceased to decrease monotonically (obviously, the validation set records were never used as training patterns). The second strategy was based on the random selection of a subset of training patterns at each epoch during the training procedure. This way it was not possible for the ANN to be influenced by the order in which the training patterns were submitted (thus possibly memorizing them). Finally, white noise in the [-0.01,0.01] range was added to each input, i.e. predictive variable. Such a small random perturbation of the input values, also known as *jittering*, favored the generalization of an ANN model because the latter learned how to associate each output pattern with a set of input intervals rather than with a single input pattern (Györgyi 1990).

The accuracy of the ANN predictions was expressed by the percentage of Correctly Classified Instances (CCI), while the significance of the deviation of the ANN predictions from a random model was tested by means of the K statistics (Cohen 1960; Fielding and Bell 1997). Details about the computation of CCI percentage and K statistics are provided in the Appendix.

Model selection

A few different basic options are available for developing models of species distribution using ANNs. The first option is to train a different model for each species, another is to train a single model that is able to simultaneously predict the distribution of all the species. A further option is to split the species list into two or more subsets on the basis, e.g. of trophic characteristics, and to train a model for each subset. In the latter case, however, the number of possible models is very high and selecting the best combination is not a straightforward task.

If only the first two options are considered, the selection of the best approach may be based on empirical tests, but there are also some theoretical considerations that should be taken into account.

In fact, when modeling the distribution of a complex set of species, such as a fish assemblage, an ANN model that predicts more than a single species is able to learn not only

the distribution of each species, but also some information about interactions among species. Of course, ecologists know that this kind of information is relevant, but in many cases their theoretical knowledge about species interactions is not adequate, as it is often based on hypotheses, personal observations, etc. Therefore, it is not easy to exploit such knowledge in modeling applications using conventional statistical methods (e.g. logistic regression). Since ANNs are able to learn from data, they are also able to learn by themselves what is relevant in species interactions and this may enhance their predictive ability.

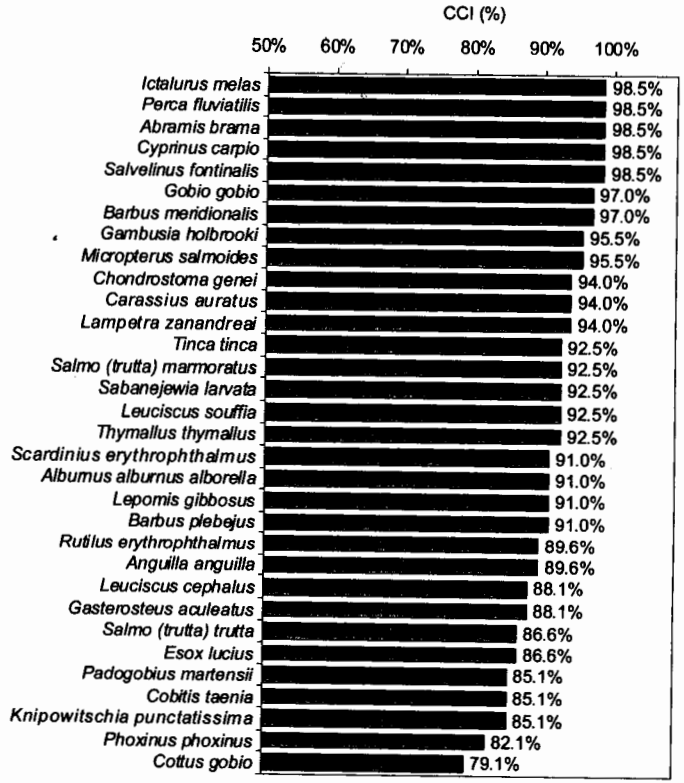


Figure 3.8.2 Percentages of Correctly Classified Instances (CCI) for the 32 modeled species. Species are sorted in descending CCI order.

Given a species assemblage containing s species, 2^s different combinations of species presence and absence data exist. In the case of our data set, 2^{32}=4 294 967 296 different patterns are theoretically possible, but only 131 different patterns were actually found in 264 observations. This is clear evidence for the non-independence of different species responses to environmental factors and for the role that biotic interactions play.

Even though simultaneously modeling all the species in a community or in an assemblage is theoretically more efficient, there are practical constraints that may hinder this approach. In fact, the complexity of the ANN structure grows very rapidly with the number of species to be modeled, and the need for training data grows proportionally. Moreover, the

set of predictive environmental variables used by the model might be more relevant to some species than to others, and this would impair the model response. In the case of fish assemblages, however, the overall number of species is usually not too large and the species response to environmental variables is rather homogeneous. Therefore, a single model approach was selected in our study.

A conventional training procedure

The first attempt at modeling the fish assemblage was based on a very conventional ANN approach, as a 20-17-32 multilayer perceptron was trained using an ordinary error back-propagation algorithm. This ANN was able to predict the presence of all the species on the basis of environmental variables. The output values it returned ranged in the [0,1] interval and therefore they could be regarded as the probability for each species of being observed. The predicted fish assemblage composition was then obtained by setting a 0.5 threshold for each output, thus converting the continuous output values into binary values (i.e. species presence or absence estimates) by means of a process that is closely related to defuzzyfication.

The overall accuracy of the ANN model was very good, as the CCI ranged from 98.5% to 79.1% (Fig. 3.8.2), while the average percentage of CCI was 91.6%. The percentage of CCI, although very convenient and easy to compute, is sometimes a misleading criterion for evaluating the ability of a model to predict species composition. In fact, it would be really appropriate if the number of presence records for a given species were exactly the same as the number of absence records, and it would still be acceptable if the ratio between presence and absence records was not too far from one. On the contrary, when the ratio becomes too small (or too large), an ANN model can be easily affected by a significant bias. For instance, when very rare species are modeled, an ANN that always returns null outputs can easily provide a very high CCI percentage. In other words, if a species were present in 2 out of 100 records (i.e. if its frequency were 2%), an ANN would be very easily able to provide 98% of CCI by constantly predicting the absence of that species. Needless to say, notwithstanding a very high CCI percentage, such an ANN could not be considered as a true model.

Therefore, another procedure was selected for evaluating the accuracy of the ANN model in the light of the actual frequency of presence or absence record for each species. In particular, the K statistics system (Cohen 1960; Fielding and Bell 1997) was applied to test whether the predictions for each species were significantly different from those of a random model or not. The ANN model was able to effectively predict 20 species out of 32, i.e. in 20 cases the K statistics was significantly different from zero (p=0.95), whereas it failed in the remaining cases (table 3.8.3).

It was evident, however, that the ability of the ANN to predict species presence and absence was strictly related to species frequency. In fact, the maximum frequency among the 12 species with non-significant K statistics was 8.71%, and 10 of them had frequencies lower than 5%. Thus, the model failed to predict several rare species, while it was quite accurate in predicting more frequent species (Fig. 3.8.3).

This result, of course, was not surprising. An ANN learns from examples, and it is obvious that it cannot learn how to correctly predict the presence of a species if the latter is only present in a few records. In these cases no ANN, or any other model, can associate the species response to patterns in the variation of predictive variables. Obviously, exactly the same problem would occur if a model were trying to predict an almost ubiquitous species.

The lack of information about the distribution of rare species is usually related to the way data are collected. In many cases the sampling effort is evenly distributed over the studied region (e.g. a river basin), because the main purpose of the sampling is the characterization of the fish assemblage composition. Therefore, stenotopic species are only found in a limited number of samples and not enough data are available about their relationships with environmental variables. A similar problem would also arise for really ubiquitous species, although in practice it is not common that a species is present in almost all the records in a data set. Moreover, density and population structure data usually provide useful hints about the environmental gradients that play a role in defining the distribution of ubiquitous species. As far as assemblage composition modeling is concerned, however, the practical effects of the lack of information about the relationships between environmental variables and species *absence* are exactly the same as those of the lack of information about the relationships between environmental variables and species *presence*.

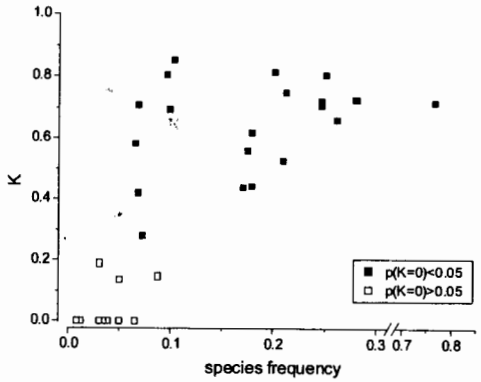


Figure 3.8.3 Conventional ANN model: K statistics vs. species frequency. The model is not reliable as far as rare species are concerned, whereas it works much better with more frequent species.

Problems in error computation

Even though no modeling technique can actually fill the gaps in the available information, it is certainly possible to improve a model by exploiting that information in a more effective way.

A conventional ANN training procedure is driven by the minimization of the Mean Square Error (MSE). As soon as the MSE becomes smaller than a previously defined value, the training procedure is stopped, assuming that the agreement between ANN output values and target (i.e. known) values is good enough. The early stopping procedure that was used in this study involves a similar role of the MSE, although the latter is minimized with respect to a validation data set that is independent of the training data set. In particular, the MSE is computed by comparing the continuous ANN outputs with the binary target values.

This approach makes perfect sense when continuous quantitative variables are involved (e.g. biomass, concentration, etc.), but it is not adequate when species composition is taken into account. There are at least three reasons for this inadequacy and they are probably not as obvious at they should be.

Table 3.8.3 Conventional ANN model: observed and predicted frequency by species (sorted in descending order of observed frequency) and K statistics (significant values are marked with asterisks).

	observed frequency	predicted frequency	K	
Salmo (trutta) trutta	76.5%	83.3%	0.719	*
Leuciscus cephalus	28.0%	31.1%	0.727	*
Padogobius martensii	26.1%	36.4%	0.660	*
Scardinius erythrophthalmus	25.0%	28.0%	0.806	*
Esox lucius	24.6%	31.1%	0.709	*
Rutilus erythrophthalmus	24.6%	26.9%	0.723	*
Alburnus alburnus alborella	21.2%	25.8%	0.748	*
Cottus gobio	20.8%	19.3%	0.528	*
Tinca tinca	20.1%	25.0%	0.816	*
Cobitis taenia	17.8%	15.5%	0.619	*
Phoxinus phoxinus	17.8%	11.4%	0.442	*
Anguilla anguilla	17.4%	12.9%	0.560	*
Knipowitschia punctatissima	17.0%	12.1%	0.440	*
Salmo (trutta) marmoratus	10.2%	9.8%	0.853	*
Sabanejewia larvata	9.8%	11.0%	0.696	*
Ictalurus melas	9.5%	12.5%	0.807	*
Lepomis gibbosus	8.7%	0.8%	0.148	n.s.
Barbus plebejus	7.2%	2.7%	0.280	*
Chondrostoma genei	6.8%	5.7%	0.709	*
Gasterosteus aculeatus	6.8%	6.4%	0.419	*
Carassius auratus	6.4%	0.0%	0.000	n.s.
Gobio gobio	6.4%	7.2%	0.583	*
Leuciscus souffia	4.9%	0.0%	0.000	n.s.
Thymallus thymallus	4.9%	0.4%	0.137	n.s.
Lampetra zanandreai	3.8%	0.0%	0.000	n.s.
Gambusia holbrooki	3.4%	0.0%	0.000	n.s.
Barbus meridionalis	3.0%	0.8%	0.190	n.s.
Micropterus salmoides	3.0%	0.0%	0.000	n.s.
Perca fluviatilis	1.1%	0.0%	0.000	n.s.
Abramis brama	0.8%	0.0%	0.000	n.s.
Cyprinus carpio	0.8%	0.0%	0.000	n.s.
Salvelinus fontinalis	0.8%	0.0%	0.000	n.s.

Firstly, when a threshold function is applied for discretizing the ANN outputs, the real contribution of each single error to the MSE strongly depends on the output value. For instance, if the target value for a given species is 0 (i.e. absence), a 0.495 output value would contribute $(0.495\text{-}0)^2$=0.245025 to the overall MSE, although it would result in a perfect agreement when the output value is transformed into a binary value by passing it to the threshold function (0.495<0.5 would be transformed into 0, i.,e. absence). A very similar output value, like, for instance, 0.505, would provide an almost identical contribution to the overall MSE $(0.505\text{-}0)^2$=0.255025, but it would be in disagreement with the target value after applying the threshold function (0.505>0.5 would be transformed into 1, i.,e. presence).

Secondly, the potential contribution of each modeled species to the MSE is identical and it varies between 0 and 1. Although this makes perfect sense from a computational point of view, it fails to capture the real effect of different errors in different contexts, because it does not weight each error according to its impact on the characterization of the species assemblage structure. In fact, a wrong prediction about a single species might have a limited effect on the overall composition of the predicted assemblage if the latter included many other species, while it might completely change the assemblage structure if the latter included only a few species. In other words, each species has an ecological "meaning" that depends not only on its ecological characteristics, but also on the way the species combines with other species, i.e. on the assemblage structure.

Finally, the efficiency of sampling is usually not homogenous, even within a single study. For instance, it is much more likely that a species, although present at a given site, escapes from sampling devices in a large river than in a small stream. Therefore, the contributions of different species to the error computation should not be simply added to each other, as in the case of MSE.

In conclusion, species presence and absence data are not to be used as mere numbers (i.e. as 0s and 1s) in the error computations that are needed for optimizing species composition models. As a consequence, the MSE is not an appropriate measure of the error in such models.

An enhanced training procedure

Several options exist for implementing an ecologically sound procedure for error computation, although not all the problems that were mentioned in the previous section can be solved. Since it is clear that the role of each species depends on other species, i.e. on species assemblage structure, a binary similarity coefficient may provide a simple yet effective way to measure the difference between the model outputs (predicted assemblage) and the target values (observed assemblage). This solution leads to a different problem, i.e. the selection of the most appropriate similarity coefficient. However, this is a common problem in ecological multivariate data analysis and most ecologists are acquainted with it and are certainly able to select a suitable coefficient. In our case study, we were able to assume that the fish assemblage composition was recorded very accurately at every sampling site. This implied that species absence in samples might be regarded as reliable information. Therefore, a symmetrical similarity coefficient that slightly emphasized differences in species composition was selected as a measure for model errors. In particular, the Rogers and Tanimoto (1960) similarity coefficient (S_{jk}) was chosen and transformed into a dissimilarity coefficient (D_{jk}), which was monotonically related to the error in the species composition prediction:

$$S_{jk} = \frac{a+d}{a+2b+2c+d} \qquad D_{jk} = 1 - S_{jk}$$

In the above formula a and d are the number of species whose presence (a) or absence (d) are correctly predicted, whereas b and c are the number of species present that are not predicted by the model and vice-versa.

The conventional ANN training procedure was then modified in order to use the mean dissimilarity between model outputs and validation patterns (i.e. samples) as the criterion for controlling ANN learning. In particular, the training procedure was halted as soon as the mean dissimilarity began to increase. This allowed an optimal generalization of ANN learning, which only takes place during the first part of the training procedure, i.e. while the error (the dissimilarity, in this case) decreases monotonically (Fig. 3.8.4).

The results of this enhanced training procedure were almost identical to those of the conventional procedure in terms of CCI percentages, but they showed a substantial improvement when other criteria were taken into account. In fact, while the average value for the CCI was 91.8%, i.e. only 0.2% higher than the one obtained by conventional training, the differences between predicted and observed species frequencies, as computed on the basis of the whole test set, were substantially smaller than in the case of conventional training (2.2% and 3.5% in absolute values, respectively).

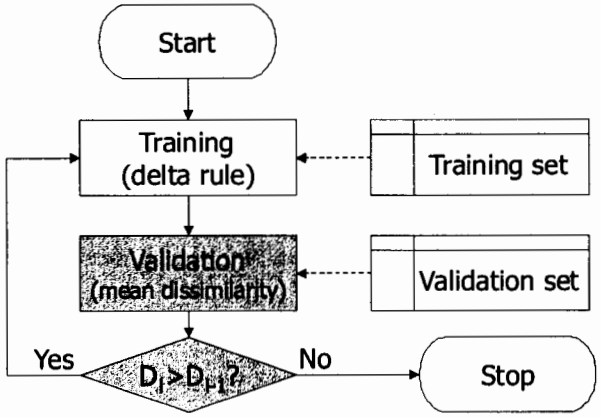


Figure 3.8.4 The training procedure for the enhanced ANN model. The modified steps are shown on grey background.

However, the most important advantage of the modified training procedure over the conventional one was in its ability to obtain better predictions for those species whose frequency was smaller than 10% (Table 3.8.4, but see also Table 3.8.3).

Moreover, the only species whose presence was never predicted by the model were the two rarest species, namely *Cyprinus carpio* and *Salvelinus fintinalis*, while the conventionally trained model was not able to predict the presence of 9 species out of 32.

Finally, the K statistics results were on average much higher than in the conventionally trained model (0.59 and 0.42, respectively), and only 5 out of the 7 less frequent species were associated to K values that were not significantly different from zero. This implied that the enhanced model was unable to predict only 5 species, while the conventionally trained model failed with 12 species.

In order to summarize the differences between the conventional (MSE-based) ANN model and the enhanced (dissimilarity-based) one, it is useful to compare the K statistics species by species, as shown in Fig. 3.8.5. The small boxes show the K values for the conventional model (solid boxes) and for the enhanced one (white boxes), while the whisker on the left of each box indicates the lower end of the confidence interval of the K statistics (the upper one is not relevant in this case, so it was omitted). Obviously, the K statistics is not significantly different from zero (at a probability level p=0.95) if the left whisker intersects the vertical axis at $K=0$. The boxes on the vertical axis with no whisker on the left show those cases in which the K statistics was not computed because the model always predicted the absence of the corresponding species. The species have been sorted according to their frequency, shown in parentheses on the right of each species name.

It is very easy to notice that there were no cases in which the conventional training provided higher K values than the enhanced model, but the most striking difference between the two models can be observed for the less frequent species. In fact, the enhanced model led to dramatic improvements in the predictive ability and in several cases the K statistics for the enhanced model was significant, while it was not significant or not even computable for the conventional model.

In the case of the enhanced model only five species were associated with values of the K statistics that were not significant, while twelve species were in that situation when the conventional model was used. It is interesting to noe that the largest changes in K values were observed for species whose frequency ranged from 3% to 9%. These species, that cannot be considered as truly rare species, are certainly associated with particular physical, chemical and biotical conditions and play a relevant role in defining the ecological characteristics of the fish assemblage.

Table 3.8.4 Enhanced ANN model: observed and predicted frequency by species (sorted in descending order of observed frequency) and K statistics (significant values are marked with an asterisk).

	observed frequency	predicted frequency	K	
Salmo (trutta) trutta	76.5%	74.6%	0.726	*
Leuciscus cephalus	28.0%	24.6%	0.805	*
Padogobius martensii	26.1%	22.0%	0.767	*
Scardinius erythrophthalmus	25.0%	23.5%	0.836	*
Esox lucius	24.6%	21.2%	0.754	*
Rutilus erythrophthalmus	24.6%	21.6%	0.765	*
Alburnus alburnus alborella	21.2%	19.7%	0.790	*
Cottus gobio	20.8%	12.5%	0.640	*
Tinca tinca	20.1%	17.4%	0.824	*
Cobitis taenia	17.8%	15.2%	0.675	*
Phoxinus phoxinus	17.8%	14.0%	0.615	*
Anguilla anguilla	17.4%	13.3%	0.721	*
Knipowitschia punctatissima	17.0%	13.6%	0.665	*
Salmo (trutta) marmoratus	10.2%	9.1%	0.876	*
Sabanejewia larvata	9.8%	8.3%	0.794	*
Ictalurus melas	9.5%	8.3%	0.829	*
Lepomis gibbosus	8.7%	2.3%	0.375	*
Barbus plebejus	7.2%	4.5%	0.603	*
Chondrostoma genei	6.8%	4.5%	0.709	*
Gasterosteus aculeatus	6.8%	3.8%	0.601	*
Carassius auratus	6.4%	1.9%	0.415	*
Gobio gobio	6.4%	4.5%	0.603	*
Leuciscus souffia	4.9%	2.3%	0.476	*
Thymallus thymallus	4.9%	1.5%	0.458	*
Lampetra zanandreai	3.8%	1.5%	0.485	*
Gambusia holbrooki	3.4%	0.4%	0.195	n.s.
Barbus meridionalis	3.0%	1.5%	0.560	*
Micropterus salmoides	3.0%	1.1%	0.490	*
Perca fluviatilis	1.1%	0.4%	0.497	n.s.
Abramis brama	0.8%	0.4%	0.394	n.s.
Cyprinus carpio	0.8%	0.0%	0.000	n.s.
Salvelinus fontinalis	0.8%	0.0%	0.000	n.s.

Conclusions

Predicting the species composition of fish assemblages on the basis of environmental descriptors is a feasible task that can be carried out either by means of conventional probabilistic models (e.g. Oberdorff et al. 2001) or by means of ANNs (e.g. Aguilar Ibarra et al. 2003; Joy and Death # 3.5; Olden and Jackson 2001). ANNs have been successfully used in these applications, as they allow exploitation of heterogeneous sources of information in a

very effective way (Scardi and Harding 1999). Moreover, ANNs may be easily enhanced and adapted to specific modeling tasks (Scardi 2001), as they are entirely empirical tools.

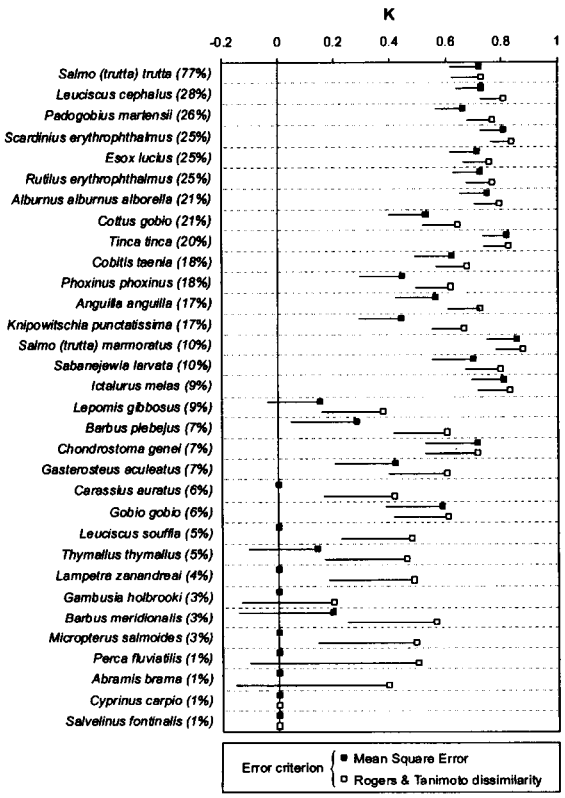


Figure 3.8.5 A comparison of K statistics values for the conventional model, using Mean Square Error as the error criterion (black squares), and the enhanced model, using Rogers and Tanimoto (1960) dissimilarity instead (white squares). The line on the left of each square shows the lower limit of the confidence interval of the K statistics. Therefore, when the line (or the symbol) intersects the vertical axis at K=0 the K statistics is not significantly different from zero (p=0.95).

Even though ANNs are the most effective tools for modeling species composition (Olden and Jackson 2002), they cannot solve problems that arise from a lack of relevant information. In fact, in many cases the only predictive variables that are readily available for the modeler are those that can be obtained from cartographic records or direct observation. Other sources of information that involve sampling and laboratory analyses are usually less abundant and therefore play a secondary role. Moreover, species distribution data are also

scarce, and distributed in space according to the local resources for monitoring activities rather than on the basis of a suitable and consistent sampling design. Therefore, predicting the species assemblage composition is not feasible without compromises. For instance, accurate ANN models can be trained on a regional scale, or focus on species assemblages simpler than communities. Our application, dealing with fish assemblages in northeastern Italian streams and rivers, belongs to this category and is certainly an example of successful modeling that can be used in practical applications. For instance, our model can be considered as a generator of expected fish assemblages, i.e. of biotic reference conditions in the light of the EU Water Framework Directive.

In particular, our model predicts the assemblage structure on the basis of environmental descriptors that are mainly (but not exclusively) focused on the geo-morphological characteristics and is based on data from real assemblages, as observed in a number of real sites. Therefore, the predicted assemblage is not just the one that is considered present at a theoretical pristine site, but a compromise that represents the more likely biotic response given a number of existing constraints, mainly related to the long term anthropogenic impacts on pristine ecosystems (e.g. changes in land usage, introduction of exotic species, modification of river banks, etc.). In regions where pristine conditions have not existed for several centuries, this is probably the only meaningful way to define reference conditions.

The ANN models presented here are not only an achievement in applied ecological research, as they also point out more general problems in species distribution modeling and provide solutions for them.

The most general scientific issue that emerged from our work is that very rare and very frequent species cannot be effectively modeled unless enough information is available. This obviously does not happen in many real studies, in which the only acceptable solution should be based on several species-specific sampling designs, i.e. on multiple sampling designs tailored to fit the distribution of each studied species.

Another relevant scientific issue that was highlighted by our work was the need for adequate error measurements in ecological applications. In fact, conventional criteria like MSE may fail when applied to data that are not strictly quantitative, like species presence and absence data. These data are binary from a formal point of view, but they cannot be treated just as sequences of 1s and 0s. Each species contributes to the assemblage structure in a way that depends simultaneously on its ecological characteristics and on the composition of the assemblage. Therefore, some errors in predicting species composition might be more relevant than others. For instance, in many upstream sites the only fish species is *Salmo trutta trutta*, which is also very frequent as a member of much more complex assemblages in other sites downstream. It is obvious that not predicting its presence in an upstream site would be a much more severe error than not predicting its presence elsewhere.

Using a binary dissimilarity coefficient instead of MSE as the criterion for measuring prediction errors provided a significant enhancement of a conventional ANN model. Even though the functioning of the error back-propagation algorithm was not changed, the modified training procedure relied on the minimization of the mean dissimilarity as a criterion for stopping the learning phase, thus allowing optimal generalization of the model. In other words, the enhanced training procedure did not change the way the ANN model learned, but it changed the conditions for stopping its optimization.

In our application the Rogers and Tanimoto (1960) dissimilarity was used, because we were confident about the reliability of our absence data and because we wanted to stress differences rather than resemblances between assemblages. In different situations, however, other coefficients would prove more adequate. For instance, if absence data are not completely reliable (e.g. because of avoidance of the catching net) an asymmetric dissimilarity that only takes into account presence data, like that based on Jaccard's coefficient (Jaccard 1908), could be more appropriate.

The enhanced training procedure not only improved the overall accuracy of the species composition predictions, but it also significantly increased the ability of the model to correctly predict the occurrence of rare species, thus mitigating the effects of the unbalanced availability of information about rare species that was previously mentioned.

In order to obtain further improvements of species composition models, however, changes in the modeling strategies should be coupled with the optimization of the sampling strategies. In fact, modeling rare or ubiquitous species is only feasible if adequate information is available, such as the ratio between the number of absence and presence records in training and validation data set which should be as close to one as possible, while the variability of the environmental descriptors within each subset, i.e. within the presence or absence subsets, should be maximum. Therefore, *ad hoc* sampling designs that significantly deviate from the usual monitoring approaches are needed. This shortcoming is not specific to ANNs, as it obviously affects any modelling technique.

The enhanced ANN model presented in this chapter was incorporated into the software tool that was published as one of the deliverables of the PAEQANN project and that can be found in the CD attached to this book. Therefore, the readers will be able to experiment the model on their own, to check its results and compare the predictions it provides with those of other models.

Appendix

Both the percentage of Correctly Classified Instances (CCI) and the *K* statistics (Cohen 1960; Fielding and Bell 1997) are based on the confusion matrix, i.e. on a 2 x 2 contingency table in which the predicted presence and absence of a taxon are compared with their observed counterpart. In particular, if each case is expressed as a proportion p_{ij}, then the confusion matrix will be:

		Predicted 1	Predicted 0
Observed	1	p_{11}	p_{12}
	0	p_{21}	p_{22}

and the sum of its elements will be 1. The CCI percentage will then be computed as:

$$CCI\% = 100 \cdot \sum_{i=1}^{2} p_{ii}$$

The *K* statistic can be easily computed from the same confusion matrix. The observed (P_o) and expected (P_e) proportion of agreement between observed and predicted data are the basis for the *K* statistics computation:

$$K = \frac{P_o - P_e}{1 - P_e}$$

In particular, P_o is closely related to CCI%, whereas P_e depends on the number of cases in all the elements of the confusion matrix:

$$P_o = \sum_{i=1}^{2} p_{ii} \qquad P_e = \sum_{i=1}^{2} \left(\sum_{j=1}^{2} p_{ij} \cdot \sum_{j=1}^{2} p_{ji} \right)$$

In order to test the significance of the deviation from zero of the *K* statistics, the standard error s_{K0} has to be computed, because the ratio between *K* and s_{K0} is distributed as the standardized normal variate *Z*. The standard error s_{K0} can be obtained as:

$$s_{K0} = \frac{\sqrt{P_e + P_e^2 - C}}{(1 - P_e) \cdot \sqrt{n}} \qquad Z = \frac{K}{s_{K0}}$$

where n is the number of cases considered in the confusion matrix and C can be obtained as

$$C = \sum_{i=1}^{2} \left[\sum_{j=1}^{2} p_{ij} \cdot \sum_{j=1}^{2} p_{ji} \cdot \left(\sum_{j=1}^{2} p_{ij} + \sum_{j=1}^{2} p_{ji} \right) \right]$$

It is very important, however, to remember that the standard error s_{K0} is not exactly the same as that needed, for instance, to compute the two-sided confidence interval for *K*.